

变集对联系度的扩充粗糙集模型及其属性约简^{*})

刘富春

(广东工业大学应用数学学院 广州 510090) (中山大学计算机系 广州 510275)

摘要 鉴于现有的不完备信息系统的扩充方法都存在一定程度的局限性,本文运用集对分析方法,引入个体之间的联系度的概念,定义一个具有自反和对称性的集对容差关系 $SP_A(\alpha, \beta)$,提出了一种变集对联系度的扩充模型。这种扩充模型是对 Kryszkiewicz 的基于容差关系的扩充模型和王国胤的基于限制容差关系的扩充模型的改进和推广,既保留了这些模型的优点,克服了它们的局限性,又可以通过对同一度 α 和对立度 β 的调节和控制,在保证不完备信息系统中个体划分的准确性的同时,增加了其灵活性。然后,给出了相应于这种模型的属性约简算法。最后,通过一个具体例子,说明这种模型对处理不完备信息系统的有效性和优越性。

关键词 不完备信息系统,集对分析方法,变集对联系度,属性约简

Variable Set-Pair Connectivity Extension of Rough Set Model and its Attributes Reduction

LIU Fu-Chun

(Faculty of Applied Mathematics, Guangdong University of Technology, Guangzhou 510090)

(Department of Computer Science, Zhongshan University, Guangzhou 510275)

Abstract Because of the limitations in the existing extensions of rough set theory, in this paper a new extension, variable set-pair connectivity rough set model, is proposed by introducing some concepts such as the connectivity between individuals and set-pair tolerance relation. It is not only an improvement of the extension based on tolerance relation, but also a generalization of the extension based on limited tolerance relation. By an example, it is verified that the new extension can reserve the advantage of the existing extensions, furthermore, it can increase the flexibility of data handling by adjusting the value of the connectivity (α, β) . Therefore, the variable set-pair connectivity rough set model is more convenient and effective than the previous extensions in practice.

Keywords Incomplete information system, Set-pair analysis, Variable set-pair connectivity, Attributes reduction

由 Pawlak 等人提出的粗糙集理论^[1],是一种新的处理不完整与不确定知识的软计算工具。它在数据挖掘、人工智能、机器学习和模式识别等领域已经取得了很大的成功^[1,3],但是,应用这种经典粗糙集理论的一个重要前提就是它的处理对象必须是完备信息系统。而在现实生中,由于测量误差、数据获取能力不足等原因,使得大量的信息系统都是不完备的,因此,需要对经典粗糙集理论进行适当的扩充^[4~7]。目前,主要有基于容差关系^[4]、相似关系^[5]、量化容差关系^[5]和限制容差关系^[6]的扩充模型,但是它们都存在一定程度的局限性^[7]。

集对分析方法^[8]是赵克勤教授近年来提出的,用于研究集合之间相互关系的一种新理论,现在它已经得到了广泛的应用^[7~9]。本文利用集对分析方法,通过引入个体之间的联系度的概念,定义一个具有自反性和对称性的集对容差关系 $SP_A(\alpha, \beta)$,提出了一种变集对联系度的扩充粗糙集模型。这种扩充模型是对 Kryszkiewicz 的基于容差关系的扩充模型^[4]和王国胤的基于限制容差关系的扩充模型^[6]的改进和推广,既保留了这些模型的优点,克服了它们的局限性,又可以通过对同一度 α 和对立度 β 的调节和控制,在保证个体划分的准确性的同时,增加了其灵活性。然后,给出了相应于这种模型的属性约简算法。最后,通过一个具体例子,说明这种模型对处理不完备信息系统的有效性和优越性。

1 变集对联系度的扩充粗糙集模型

集对分析方法^[8]是由赵克勤教授近年来提出的,用于研究集合之间相互关系的一种新理论,其核心思想是把被研究的客观事物之确定性联系和不确定性联系作为一个系统来处理。

定义 1 设 $(U, A \cup \{d\})$ 是一个不完备信息系统,其中 U 是对象集, A 是条件属性集, d 是决策属性。 $\forall x, y \in U, D \subseteq A \cup \{d\}$, 定义 x 与 y 在 D 下的集对属性集如下:

$$M(x, y) = \{a \in D \mid a(x) \neq * \wedge a(y) \neq * \wedge a(x) = a(y)\}$$

$$N(x, y) = \{a \in D \mid a(x) \neq * \wedge a(y) \neq * \wedge a(x) \neq a(y)\}$$

$$K(x, y) = \{a \in A \mid a(x) = * \vee a(y) = *\}$$

定义 2 设 $(U, A \cup \{d\})$ 是一个不完备信息系统。 $\forall x, y \in U, D \subseteq A \cup \{d\}$, 定义 x 与 y 在 D 下的集对联系度 $U_D(x, y)$ 为:

$$U_D(x, y) = s_1 + s_2 i + s_3 j$$

这里 $s_1 = |M(x, y)|/n$ 称为 x, y 在 D 下的同一度, $s_2 = |N(x, y)|/n$ 称为 x, y 在 D 下的对立度, $s_3 = |K(x, y)|/n$ 称为 x, y 在 D 下的差异度, 其中 $n = |D|$ 。

^{*}) 广东省自然科学基金项目(020146, 031541)和广东工业大学青年基金项目(042027)。刘富春 讲师,博士研究生,主要研究方向:粗糙集理论、数理逻辑。

显然, $s_1 + s_2 + s_3 = 1$

Kryszkiewicz 提出的基于容差关系 T 定义为:

定义 3^[4] 设 $(U, A \cup \{d\})$ 是一个不完备信息系统, $D \subseteq A \cup \{d\}$, 则容差关系 T 的定义为:

$$T_D = \{(x, y) \in U \times U \mid \forall a \in D (a(x) = * \vee a(y) = * \vee a(x) = a(y))\}$$

在基于容差关系的扩充模型中, 不确定值“*”实际上被认为是和所有已知属性值都相等, 因此会导致那些没有任何明确相同的已知属性值的个体, 都被认为是不可分辨而被误判在同一个容差类中^[6]。

为此, 王国胤提出了一种限制容差关系 L 。

定义 4^[6] 设 $(U, A \cup \{d\})$ 是一个不完备信息系统, $D \subseteq A \cup \{d\}$, 则限制容差关系 L 的定义为:

$$L_D = \{(x, y) \in U \times U \mid \forall a \in D (a(x) = a(y) = *)$$

$$\vee (P(x) \cap P(y) \neq \emptyset \wedge \forall a \in D ((a(x) \neq *) \wedge (a(y) \neq *) \rightarrow (a(x) = a(y))))\}.$$

其中 $P(x) = \{a \in D \mid a(x) \neq *\}$ 。

容差关系^[4]和相似关系^[6]是对不可分辨关系的扩充的两个极端情形: 容差关系的条件太宽松, 相似关系的条件又太紧, 而限制容差关系恰好是介于它们之间, 因此, 基于限制容差关系的扩充模型优于基于相似关系和基于容差关系的扩充模型, 更符合实际情况^[6]。但是它也存在一定程度的不足:

(1) 没有一个明确相同的已知属性值的个体 (如 $a = (1, *, *, \dots, *)$ 和 $b = (*, *, *, \dots, *)$), 被认为是可分辨的, 但对于所有属性值都不确定的个体 (如 $b = (*, *, *, \dots, *)$ 和 $c = (*, *, *, \dots, *)$), 却认为是不可分辨的, 这与实际情况不相符合。(2) 个体之间只要有一个属性值不同 (如 $x = (1, 2, 3, 4, \dots, 100)$ 和 $y = (2, 2, 3, 4, \dots, 100)$) 就认为是完全不同的, 必须被划分在不同的类中。这将会导致知识库的划分太细, 划分类的个数太多, 特别不适于对大型不完备信息系统的处理。(3) 利用这些模型获得知识大多都只是一个层次上的, 难以从多个级别上对原有系统进行分析处理。

鉴于以上局限性, 下面运用集对分析方法, 提出一种变集对联系度的扩充粗糙集模型。

定义 5 设 $(U, A \cup \{d\})$ 是一个不完备信息系统, $0 \leq \alpha, \beta \leq 1$, 集对 (α, β) 容差关系 $SP_A(\alpha, \beta)$ 定义为:

$$SP_A(\alpha, \beta) = \{(x, y) \in U \times U \mid U_A(x, y) = s_1 + s_2 i + s_3 j, \text{ 其中 } s_1 > \alpha \text{ 且 } s_2 \leq \beta\} \cup I.$$

这里 $I = \{(x, x) \mid x \in U\}$ 是 U 上的恒等关系。

个体 x 的集对 (α, β) 容差类为: $SP_A(\alpha, \beta)(x) = \{y \in U \mid (x, y) \in SP_A(\alpha, \beta)\}$,

对象集 X 的下近似集和上近似集定义为:

$$\underline{SP}_A(\alpha, \beta)(X) = \{x \in U \mid SP_A(\alpha, \beta)(x) \subseteq X\};$$

$$\overline{SP}_A(\alpha, \beta)(X) = \{x \in U \mid SP_A(\alpha, \beta)(x) \cap X \neq \emptyset\}.$$

显然, 集对 (α, β) 容差关系 $SP_A(\alpha, \beta)$ 具有自反性和对称性, 不具有传递性。

下面将集对 (α, β) 容差关系和容差关系、限制容差关系进行比较分析, 可以得到如下定理。

定理 1 设 $(U, A \cup \{d\})$ 是不完备信息系统, T_A 为容差关系, $SP_A(\alpha, \beta)$ 为集对 (α, β) 容差关系, 则当 $\beta = 0$ 时, $SP_A(\alpha, \beta) \subseteq T_A$ 。

证明: 对任意 $(x, y) \in SP_A(\alpha, \beta)$, 根据定义 5 得, $(x, y) \in I$, 或者 $U_A(x, y) = s_1 + s_2 i + s_3 j$, 其中 $s_1 > \alpha$ 且 $s_2 \leq \beta$ 。如果 $(x, y) \in I$, 则 $\forall a \in A$ 都有 $a(x) = a(y)$, 所以 $(x, y) \in$

T_A 。如果 $U_A(x, y) = s_1 + s_2 i + s_3 j$, 其中 $s_1 > \alpha$ 且 $s_2 \leq \beta$ 。因为 $\beta = 0$, 所以 $s_2 = 0$, 即 $N(x, y) = \{a \in A \mid a(x) \neq * \wedge a(y) \neq * \wedge a(x) \neq a(y)\} = \emptyset$, 从而 $\forall a \in A$ 都有 $a(x) = * \vee a(y) = * \vee a(x) = a(y)$, 即 $(x, y) \in T_A$ 。因此, 当 $\beta = 0$ 时, $SP_A(\alpha, \beta) \subseteq T_A$ 。

定理 2 设 $(U, A \cup \{d\})$ 是不完备信息系统, L_A 为限制容差关系, $SP_A(\alpha, \beta)$ 为集对 (α, β) 容差关系。如果 $\forall x \in U, P(x) \neq \emptyset$, 则当 $\alpha = \beta = 0$ 时, $SP_A(\alpha, \beta) = L_A$ 。

证明: 对任意 $(x, y) \in SP_A(\alpha, \beta)$, 根据定义 5 得, $(x, y) \in I$, 或者 $U_A(x, y) = s_1 + s_2 i + s_3 j$, 其中 $s_1 > \alpha$ 且 $s_2 \leq \beta$ 。所以当 $\alpha = \beta = 0$ 时, 有 $s_1 > 0, s_2 = 0$, 即 $M(x, y) = \{a \in A \mid a(x) \neq * \wedge a(y) \neq * \wedge a(x) = a(y) \neq \emptyset\}$, 但 $N(x, y) = \{a \in A \mid a(x) \neq * \wedge a(y) \neq * \wedge a(x) \neq a(y)\} = \emptyset$ 。从而 $P(x) \cap P(y) \neq \emptyset$, 而且 $\forall a \in A$ 都有 $(a(x) \neq * \wedge a(y) \neq *) \rightarrow (a(x) = a(y))$ 成立, 故 $(x, y) \in L_A$ 。反之, 对任意 $(x, y) \in L_A$, 因为 $\forall x \in U, P(x) \neq \emptyset$, 所以 $P(x) \cap P(y) \neq \emptyset$, 而且 $\forall a \in A$ 都有 $(a(x) \neq * \wedge a(y) \neq *) \rightarrow (a(x) = a(y))$ 。于是 $M(x, y) \neq \emptyset$, 且 $N(x, y) = \emptyset$, 即 $s_1 > 0, s_2 = 0$ 。从而 $(x, y) \in SP_A(\alpha, \beta)$, 其中 $\alpha = \beta = 0$ 。这就证明了 $SP_A(\alpha, \beta) = L_A$ 。

注记 1: 文[6]指出, 容差关系 T 的条件太宽松, 是对经典粗糙集理论中不可分辨关系的扩充的极端情形, 而定理 1 表明, 当 $\beta = 0$ 时, 集对 (α, β) 容差关系正好加强了容差关系的条件。而定理 2 表明, 对于满足 $\forall x \in U, P(x) \neq \emptyset$ 的不完备信息系统 $(U, A \cup \{d\})$ 来说, 当 $\alpha = \beta = 0$ 时, 集对 (α, β) 容差关系与限制容差关系相同。因此, 变集对联系度的扩充粗糙集模型是对 Kryszkiewicz 提出的基于容差关系的扩充模型^[4]和王国胤提出的基于限制容差关系的扩充模型^[6]的推广和改进, 既保留了这些模型的优点, 克服了它们的局限性, 又可以通过对同一度 α 和对立度 β 的调节和控制, 在保证个体划分的准确性的同时, 增加了其灵活性, 对不完备信息系统中模糊和不确定性知识的处理更加合理、有效, 特别适于对大型不完备信息系统的处理。

2 变集对联系度的扩充粗糙集模型中的属性约简

属性约简是粗糙集理论的核心内容之一, 其目的是在保持知识库的分类或决策能力不变的前提下, 删除决策表中冗余或不重要的属性。下面根据集对 (α, β) 容差关系 $SP_A(\alpha, \beta)$ 的定义, 给出近似约简的概念。

定义 6 设 $(U, A \cup \{d\})$ 是不完备信息系统, $0 \leq \alpha, \beta \leq 1$, $SP_A(\alpha, \beta)$ 是集对 (α, β) 容差关系。假设对于条件属性 $a \in A$, 记

$$SP_{A-\{a\}}(\alpha, \beta) = \{(x, y) \in U \times U \mid U_{A-\{a\}}(x, y) = s_1 + s_2 i + s_3 j, \text{ 其中 } s_1 > \alpha \text{ 且 } s_2 \leq \beta\} \cup I.$$

如果 $SP_A(\alpha, \beta) = SP_{A-\{a\}}(\alpha, \beta)$, 则称属性 a 在集对 (α, β) 容差关系下是不必要的; 否则, 称属性 a 是必要的。

条件属性子集 $D = \{a \in A \mid a \text{ 在集对 } (\alpha, \beta) \text{ 容差关系下是必要的}\}$ 称为 A 的 (α, β) 近似约简。

根据定义 6, 我们可以给出变集对联系度的扩充粗糙集模型的属性约简算法。设 $(U, A \cup \{d\})$ 是一个不完备信息系统, 算法是以条件属性集 A 为起点, 按照广度优先搜索策略, 依次从条件属性集中删除不必要的属性, 直到求出 A 的一个 (α, β) 近似约简为止。

算法 求条件属性集 A 的 (α, β) 近似约简 Red 。

输入: 一个不完备信息系统 $(U, A \cup \{d\})$; 同一度 α ; 对立

度 β 。
 输出: A 的一个 (α, β) 近似约简 Red 。
 步骤 1. 令 $Red = A$, 计算 U 上的集对 (α, β) 容差关系 $SP_{Red}(\alpha, \beta)$ 。
 步骤 2. 对每个 $a \in Red$, 计算 U 上的集对 (α, β) 容差关系 $SP_{Red - \{a\}}(\alpha, \beta)$ 。
 步骤 3. 如果存在某个 $a \in Red$, 使得 $SP_{Red - \{a\}}(\alpha, \beta) = SP_{Red}(\alpha, \beta)$, 则取 $Red = Red - \{a\}$, 转向步骤 2; 否则, 转向步骤 4。
 步骤 4. 输出 Red , 它就是 A 的一个 (α, β) 近似约简 Red 。

记注 2: 本算法的优点是, 它可以通过对同一度 α 和对立度 β 的取值进行控制和调节, 按照广度优先搜索策略, 得到条件属性集 A 的一个满足一定同一度和对立度要求的近似约简。因此, 它在保证属性约简的准确性的同时增加了其灵活性, 更适于对大型不完备信息系统的处理。

3 实例分析

为考察本文提出的变集对联系度的扩充粗糙集模型及其属性约简算法的有效性和和优越性, 下面用一个实际的不完备信息系统对于同一度 $\alpha = 0.2$ 和对立度 $\beta = 0.1$ 的情形进行分析。
 设 $(U, A \cup \{d\})$ 是一个不完备信息系统, 其中对象集 $U = \{x_1, x_2, \dots, x_{12}\}$, 条件属性集 $A = \{a_1, a_2, a_3, a_4\}$, d 是决策属性^[2]。

表 1 不完备信息系统 $(U, A \cup \{d\})$

对象集 U	条件属性集 A				决策属性 d
	a ₁	a ₂	a ₃	a ₄	
x ₁	3	2	1	0	Φ
x ₂	2	3	2	0	Φ
x ₃	2	3	2	0	Ψ
x ₄	*	2	*	1	Φ
x ₅	*	2	*	1	Ψ
x ₆	2	3	2	1	Ψ
x ₇	3	*	*	3	Φ
x ₈	*	0	0	*	Ψ
x ₉	3	2	1	3	Ψ
x ₁₀	1	*	*	*	Φ
x ₁₁	*	2	*	*	Ψ
x ₁₂	3	2	1	*	Φ

步骤 1. 计算 U 上的集对 (α, β) 容差关系为:
 $SP_A(\alpha, \beta) = \{(x, y) \in U \times U \mid U_A(x, y) = s_1 + s_2 i + s_3 j, \text{ 其中 } s_1 > 0.2 \text{ 且 } s_2 \leq 0.1\} \cup I = R \cup R^{-1} \cup I$, 其中 $R = \{(x_1, x_{11}), (x_1, x_{12}), (x_2, x_3), (x_4, x_5), (x_4, x_{11}), (x_4, x_{12}), (x_5, x_{11}), (x_5, x_{12}), (x_7, x_9), (x_7, x_{12}), (x_9, x_{11}), (x_9, x_{12}), (x_{11}, x_{12})\}$, R^{-1} 是指 R 的逆关系, I 是 U 上的恒等关系。
 步骤 2. 对每个 $a \in A$, 计算 U 上的集对 (α, β) 容差关系 $SP_{A - \{a\}}(\alpha, \beta)$ 为:
 (1) $SP_{A - \{a_1\}}(\alpha, \beta) = R_1 \cup R_1^{-1} \cup I$, 其中 $R_1 = \{(x_1, x_{11}), (x_1, x_{12}), (x_2, x_3), (x_4, x_5), (x_4, x_{11}), (x_4, x_{12}), (x_5, x_{11}), (x_5, x_{12}), (x_7, x_9), (x_9, x_{11}), (x_9, x_{12}), (x_{11}, x_{12})\}$ 。
 (2) $SP_{A - \{a_2\}}(\alpha, \beta) = R_2 \cup R_2^{-1} \cup I$, 其中 $R_2 = \{(x_1, x_{12}), (x_2, x_3), (x_4, x_5), (x_7, x_9), (x_7, x_{12}), (x_9, x_{12})\}$ 。

(3) $SP_{A - \{a_3\}}(\alpha, \beta) = R_3 \cup R_3^{-1} \cup I$, 其中 $R_3 = \{(x_1, x_{11}), (x_1, x_{12}), (x_2, x_3), (x_4, x_5), (x_4, x_{11}), (x_4, x_{12}), (x_5, x_{11}), (x_5, x_{12}), (x_7, x_9), (x_7, x_{12}), (x_9, x_{11}), (x_9, x_{12}), (x_{11}, x_{12})\}$ 。
 (4) $SP_{A - \{a_4\}}(\alpha, \beta) = R_4 \cup R_4^{-1} \cup I$, 其中 $R_4 = \{(x_1, x_4), (x_1, x_5), (x_1, x_7), (x_1, x_9), (x_1, x_{11}), (x_1, x_{12}), (x_2, x_3), (x_2, x_6), (x_3, x_6), (x_4, x_5), (x_4, x_9), (x_4, x_{11}), (x_4, x_{12}), (x_5, x_9), (x_5, x_{11}), (x_5, x_{12}), (x_7, x_9), (x_7, x_{12}), (x_9, x_{11}), (x_9, x_{12}), (x_{11}, x_{12})\}$ 。

步骤 3. 因为 $SP_{A - \{a_3\}}(\alpha, \beta) = SP_A(\alpha, \beta)$, 所以条件属性 a_3 在集对 (α, β) 容差关系 $(\alpha = 0.2, \beta = 0.1)$ 下是不必要的。
 步骤 4. 类似步骤 2 的方法, 计算 U 上的集对 (α, β) 容差关系 $SP_{A - \{a_3, a_1\}}(\alpha, \beta)$, $SP_{A - \{a_3, a_2\}}(\alpha, \beta)$, $SP_{A - \{a_3, a_4\}}(\alpha, \beta)$, 不难发现它们都与 $SP_{A - \{a_3\}}(\alpha, \beta)$ 不相等。
 步骤 5. 根据步骤 3 和步骤 4 得, 当 $\alpha = 0.2, \beta = 0.1$ 时, A 的一个 (α, β) 近似约简为 $Red = \{a_1, a_2, a_4\}$, 这与文[2]的结果相同。

在实际中, 特别是在处理大型不完备信息系统时, 我们可以通过改变同一度 α 和对立度 β 的取值, 得到各种不同要求下的近似约简。因此, 本文提出的变集对联系度的扩充粗糙集模型, 既保留了其它扩充模型的优点, 克服了它们的局限性, 又可以通过对同一度 α 和对立度 β 的调节和控制, 在保证个体划分的准确性的同时, 增加了其灵活性, 特别适用于对大型不完备信息系统的处理。

结束语 本文运用集对分析方法, 引入个体之间的联系度的概念, 定义一个具有自反和对称性的集对容差关系 $SP_A(\alpha, \beta)$, 提出了一种变集对联系度的扩充模型。这种扩充模型既保留了其它扩充模型的优点, 克服了它们的局限性, 又可以通过对同一度 α 和对立度 β 的调节和控制, 在保证不完备信息系统中个体划分的准确性的同时, 增加了其灵活性。然后, 给出了相应于这种模型的属性约简算法。最后, 通过一个具体例子, 说明这种模型对处理不完备信息系统的有效性和优越性。

参 考 文 献

- 1 Pawlak Z, Busse J G, et al. Rough sets. Communications of the ACM [J], 1995, 38(11): 89~95
- 2 王国胤. Rough 理论与知识获取[M]. 西安: 西安交通大学出版社, 2001
- 3 张文修, 吴伟志, 梁吉业, 李德玉. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001
- 4 Kryszkiewicz M. Rough set approach to incomplete information system. Information Sciences [J], 1998, 112: 39~49
- 5 Stefanowski J, Tsoukias A. On the extension of rough sets under incomplete information. In: S Zhong, A Skowron, S Ohsuga, eds. Proc of the 7th Int'l Workshop on New Directions in Rough Sets, Data Mining, and Granular Soft Computing [M]. Berlin: Springer-Verlag, 1999. 73~81
- 6 王国胤. Rough 集理论在不完备信息系统中的扩充[J]. 计算机研究与发展, 2002, 39(10): 1238~1243
- 7 刘富春. 基于限制容差关系的集对粗糙集模型[J]. 计算机科学, 2005, 32(6)
- 8 赵克勤. 集对分析及其初步应用[M]. 浙江科学出版社, 2000
- 9 刘富春. 基于集对分析的变精度粗糙集模型[J]. 计算机工程与应用, 2005, 41(10): 74~76